学校编码：10384　　分类号_____密级_____
学号：15420111151910　　UDC_____

硕　士　学　位　论　文

分类数据中高维列联表可压缩性研究

Compressibility Research on High Way Contingency Table of the Categorical Data

孙红艳

指导教师姓名：钱 争 鸣 教授

专　业　名　称：统 计 学

论文提交日期：2014 年 3 月　日

论文答辩时间：2014 年 5 月

学位授予日期：2014 年

答辩委员会主席：_____

评　　阅　　人：_____

2014 年 3 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

2014 年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（ √ ）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

2014 年　　月　　日

摘 要

分类数据的统计分析方法是分析名义数据和有序数据的重要工具，在分类数据分析中，用列联表对数据进行分析是一种常用、直观的方法，例如，医学研究者按年龄和性别对病例进行分类建立列联表；教育工作研究人员按年龄、性别和家庭背景对学生进行分类建立列联表；经济研究者按照行业、地区、初始投资对企业成败进行分类建立列联表；市场研究者按年龄、性别和对商品的消费倾向进行分类建立列联表等。

传统的分类数据分析方法主要是对列联表进行独立性检验，随着对数线性模型的提出以及广泛应用，使得分类数据分析方法经常用于分析高维列联表，但是国内外文献中缺少对高维列联表的详细分析方法。由于高维列联表数据资料的复杂性，在分析高维列联表的时候为了更好地分析数据中变量的相关性，需要通过一些方式对列联表进行降维，也即对列联表中变量进行压缩，但不合理的压缩会导致辛普森悖论、虚假相关、虚假独立三种现象的产生，这就增大了分析列联表的难度，所以研究列联表可压缩性的方法非常重要，国内外学者对三维列联表已经有些研究，但仍缺少对高维列联表的可压缩性方面的研究。

本文通过基于交互作用与互信息、信息熵三种角度对列联表的可压缩性进行分析研究，深入探讨高维列联表可压缩的条件和实现途径，研究发现：

1、对于三维列联表只要满足变量之间存在条件独立列联表就可压缩，但对于四维列联表，尽管变量之间存在条件独立并不能保证列联表可压缩；

2、基于交互作用的对数线性模型与基于互信息的线性信息模型之间存在等价条件，两种模型分析的结果可以互相利用；

3、给出了线性信息模型设定条件变量与不设定条件变量的模型选择方法，发现所拟合的线性信息模型比对数线性模型更加简洁，在交互作用下的模型显示不可压缩，但在互信息下的模型显示可以压缩；

4、给出了基于互信息和信息熵列联表变量可压缩的方法，发现基于互信息的可压缩性方法是在考虑了变量相关性的角度对列联表进行的压缩，在压缩过程中允许损失部分不显著的相关信息；基于信息熵的可压缩性方法是在考虑变量含有不确定信息的多少而对列联表进行的压缩，在压缩的过程中不允许损失变量的任何信息；

5、给出了两种分别基于互信息和信息熵对列联表变量重要性的排序方法，发现从列联表可压缩性的角度，基于互信息的变量重要性排序方法更加准确。而从变量含有的不确定信息多少的角度，基于信息熵的变量重要性排序方法更加准确。

研究的成果对分类数据分析方法的研究深入发展做出新的贡献，对高维列联表的可压缩性方法提供了一些重要可实现的途径。

关键词：列联表压缩；辛普森悖论；交互作用；互信息；信息熵

Abstract

Categorical data analysis is an important tool for data analysis of nominal data and ordered data. In categorical data analysis, contingency table analysis of data is a common and intuitive method. For example, medical researchers categorize cases by age and gender to establish contingency table; education workers classify students by age, gender and family background to establish contingency table; economic researchers classify the success or failure of the enterprises by industry, area and initial investment to establish contingency table; market researchers categorize cases according to age, gender and the tendency of commodity consumption to establish contingency table, etc.

The traditional categorical data analysis method is mainly for the independence test of contingency table. As the log-linear model is put forward and widely used, categorical data analysis method is often used to analyze the high way contingency table, but the literature in domestic and overseas lack of detailed analysis of the high way contingency table. Because of the complexity of the high way contingency table data, we wish to find some ways to reduce the dimension of the contingency table in order to better analyze the correlation of variables, that is to compress variables of the contingency table. But unreasonable compression can lead to the Simpson Paradox, spurious correlation and spurious independence, this will increase the difficulty of contingency table analysis. So the study on compressibility method of contingency table is very important. At present , scholars in domestic and overseas have some research on the compressibility of three way contingency table, but lack of the compressibility of high way contingency table.

This paper will discuss the compressibility of high way contingency table based on the interaction, mutual information and information entropy. Research findings:

1. three way contingency table can be compressed as long as there are conditional independence between the variables, but for the four way contingency table conditional independence between the variables is not guarantee that the contingency table can be compressed;

2. There are equivalent conditions between the log-linear model based on the interaction and the linear information model based on mutual information, so the

analysis results of the two model can be used with each other;

3. The model selection methods for linear information model setting condition variable and linear information model not setting condition variable are proposed, find that the fitting results of linear information model is more concise than the linear information model. For the same example, the results of log-linear model show that the data is incompressible, but the results of linear information model show that the data can be compressed;

4. The compressible methods of contingency table based on mutual information and information entropy are proposed, find that the compressibility method based on mutual information considers the variables' correlation, in the process of compression allows loss of no significant relevant information, while the compressibility method based on information entropy considers the uncertain information of the variables, in the process of compression does not allow any information loss;

5. Two kinds of methods rank the variable importance based on mutual information and the information entropy are proposed, find that the method based on mutual information is more accurate from the angle of contingency table compressibility, but the method based on information entropy is more accurate from the angle of the uncertain information of the variables.

The research offers some ideas for the compressibility of high way contingency table and has certain contribution to the development of categorical data analysis.

Key words： The compression of contingency table ; Simpson's paradox ; Interaction; Mutual information ; Information entropy

目　录

第 1 章 绪论 1

1.1 研究背景与理论基础 1

1.1.1 研究背景 1

1.1.2 理论基础 3

1.2 国内外文献综述 7

1.2.1 国外研究成果 7

1.2.2 国内研究成果 8

1.2.3 文献综述小结 9

1.3 研究内容与创新点 10

1.3.1 研究内容 10

1.3.2 创新点 11

1.4 文章结构与论文框架图 12

1.4.1 文章结构 12

1.4.2 论文框架图 13

第 2 章 列联表中三种现象 14

2.1 辛普森悖论 14

2.2 虚假相关 15

2.3 虚假独立 17

2.4 本章小结 19

第 3 章 对数线性模型类型与独立性分析 20

3.1 二维对数线性模型类型与独立性分析 21

3.2 三维对数线性模型类型与独立性分析 21

3.3 四维对数线性模型类型与独立性分析 25

3.3.1 四维对数线性模型类型分析 26

3.3.2 四维对数线性模型独立性分析 28

3.4 本章小结 36

第 4 章 基于交互作用的列联表可压缩性分析 38

4.1 基于条件列联表中交互作用的可压缩性分析 38

4.2 基于边际列联表交互作用的列联表可压缩性分析.....41
4.2.1 二维列联表可压缩性分析.....41
4.2.2 三维列联表可压缩性分析.....42
4.2.3 四维列联表可压缩性分析.....47
4.3 本章小结.....67
第 5 章 基于互信息、信息熵的列联表可压缩性分析.....69
5.1 基于互信息的可压缩性分析.....70
5.1.1 二维线性信息模型与可压缩性分析.....70
5.1.2 三维线性信息模型与可压缩性分析.....72
5.1.3 四维线性信息模型与可压缩性分析.....82
5.2 基于信息熵的列联表可压缩性分析.....102
5.2.1 基于信息熵的列联表变量压缩方法.....102
5.2.2 基于信息熵变量的重要性分析.....107
5.3 本章小结.....110
第 6 章 结论与展望.....111
6.1 研究结论.....111
6.2 研究展望.....112
参考文献.....114
致 谢.....117

Contents

Chapter 1 Introduction.....1

1.1 Resarch Background and Theoretical Basis 1

1.1.1 Resarch Background 1

1.1.2 Theoretical Basis.....3

1.2 Literature Review.....7

1.2.1 Foreign Research Review 7

1.2.2 Domestic Research Review 8

1.2.3 Summary of Literature Review.....9

1.3 Research Contents and Possible Innovations.....10

1.3.1 Research Contents.....10

1.3.2 Possible Innovations 11

1.4 Structure of the Article and Frame Diagram.....12

1.4.1 Structure of the Article.....12

1.4.2 Frame Diagram of the Article 13

Chapter 2 Three Phenomenons in the Contingency Table.....14

2.1 Simpson's Paradox 14

2.2 Spurious Correlation 15

2.3 Spurious Independence 17

2.4 Summary of This Chapter.....19

Chapter 3 Type and Independence Analysis of the Log-linear Model 20

3.1 Type and Independence Analysis of the Two Way Log-linear Model ...21

3.2 Type and Independence Analysis of the Three Way Log-linear Model.21

3.3 Type and Independence Analysis of the Four Way Log-linear Model...25

3.3.1 Type Analysis of the Four Way Log-linear Model.....26

3.3.2 Independence Analysis of the Four Way Log-linear Model 28

3.4 Summary of This Chapter.....36

Chapter 4 Compressibility Analysis of the Contingency Table Based on the Interaction.....38

4.1 Compressibility Analysis Based on the Interaction of Conditional Contingency Table 38
4.2 Compressibility Analysis Based on the Interaction of Marginal Contingency Table 41
4.2.1 Compressibility Analysis of the Two Way Contingency Table... 41
4.2.2 Compressibility Analysis of the Three Way Contingency Table 42
4.2.3 Compressibility Analysis of the Four Way Contingency Table .. 47
4.3 Summary of This Chapter 67
Chapter 5 Compressibility Analysis of the Contingency Table Based on Mutual Information and Information Entropy 69
5.1 Compressibility Analysis Based on Mutual Information 70
5.1.1 Two Way Linear Information Model and Analysis of Compressibility 70
5.1.2 Three Way Linear Information Model and Analysis of Compressibility 72
5.1.3 Four Way Linear Information Model and Analysis of Compressibility 82
5.2 Compressibility Analysis Based on Information Entropy 102
5.2.1 The Method of Compression Based on Entropy 102
5.2.2 The Importance Analysis of Variables Based on Entropy 107
5.3 Summary of This Chapter 110
Chapter 6 Conclusion and Future Studies 111
6.1 The Research Conclusiuon 111
6.2 Future Studies 112
References 114
Acknowledgements 117

第 1 章 绪论

1.1 研究背景与理论基础

1.1.1 研究背景

在统计分析中通常面临的数据资料可以分为下面四类：

（1）计量数据。例如一个企业的资产、负债、所有者权益、利润等，一个人的身高、体重、血压等，气象学上的温度、相对湿度等，这类数据资料的特点是：原则上它的取值可以为某一区间上的任一个实数，一般来说这类资料是连续的。我们使用连续随机变量的分布对这类资料进行统计分析。

（2）计数数据。例如一个企业职工的总人数，一段时间内某个路口通过的车辆总数，一个社区居住的居民总数等，这种数据资料的特点是：通常它们的取值为非负的整数，这类资料称为计数数据。一般使用离散的随机变量的分布对这类资料进行统计分析，一般研究涉及这类内容的比较有限。

（3）名义数据。例如一个企业的组织机构代码编号，学校里的每个学生都有的学号，图书馆对每个书籍的编号，这类资料既不是计量数据，也不是计数数据，编号只是起到一个名义的作用，称这类数据资料为名义数据。

（4）有序数据。例如判断一个企业规模的大小，评定一种酒或茶的好坏，一个酒店的顾客对酒店的服务满意程度，这类数据也既不是计量数据，也不是计数数据，同时也不是名义数据，只能判断出一个顺序，而不能量化，我们称这类数据为有序数据。

这四类数据资料可以分成定量数据和定性数据，定量数据：包括计量数据和计数数据；定性数据：包括名义数据和有序数据，其中定量数据中的离散数据与定性数据一起称为分类数据。关于定量数据的研究方法已经有很多文献和书籍，但关于分类数据的研究方法则相对较少。

分类数据的统计分析方法与数理统计分析方法同样有着 100 多年的历史，但分类数据统计分析方法的发展相对连续数据的统计分析方法要慢很多，分类数据也称属性数据（包括定性数据、离散数据），是社会科学研究的组成部分，分类数据分析的应用范围非常广泛，如在市场营销学领域：研究产品的的价格是否影响产品销量；在政治学领域：分析不同类别的人不同的派别倾向；在医学领域：

研究不同药物剂量对病人疾病治疗的效果；在卫生领域：研究人们对环境保护的看法。

在社会调查过程中，常常需要遇到分类数据，分类数据是社会调查中较难分析处理的一种数据。为了进行统计分析，我们经常对分类数据的变量进行赋值，这种方法首先是由日本统计学家林知己夫提出来的。对于名义变量，即变量是没有顺序区别的不同状态，给变量值赋予不同的代号，例如，年龄用“老”和“少”作为其取值，也可以用数字“1”和“2”进行赋值，但这两个数字仅仅是一个代码，不能进行类似数值的加减运算。有序变量的取值是有严格顺序的，例如，一个人的身高可以分为低、中、高等，和名义变量相同，也可以用代表顺序的数字给它们赋予不同的数字，也不能进行加减运算。有时在进行社会调查研究时也会将一些定量数据进行分组转换成分类变量进行分析，从这个角度来看，所有的社会调查研究的原始数据，都可以看成是分类数据，对分类数据分析方法的讨论和研究有着重要的意义。

在分类数据分析中，用列联表对数据进行分析是一种常用、直观的方法，例如，医学研究者按年龄和性别对病例进行分类建立列联表；教育工作研究人员按年龄、性别和家庭背景对学生进行分类建立列联表；经济研究者按照行业、地区、初始投资对企业成败进行分类建立列联表；市场研究者按年龄、性别和对商品的消费倾向进行分类建立列联表等，一般来说三维或三维以上的列联表称为高维列联表。

目前，分类数据分析的统计分析方法在数据分析中逐渐占据着越来越重要的地位，使用列联表对分类数据进行统计分析越来越受欢迎，特别是在医学、生物学和社会学领域。这在一定程度上可以看出分类数据分析方法的快速发展，同时也可以看出列联表在分类数据分析中的重要性，很多统计研究学者已经意识到，用连续数据的统计分析方法去分析分类数据是不可取的。随着计算机的高速发展，分类数据的统计分析方法也为我们进行数据分析的时候提出了一种很好的思路，特别是高维列联表的统计分析方法越来越多地被研究人员使用。在对列联表数据的分析中，必须采用一种适合数据的自身特点的方法进行分析，否则有可能导致辛普森悖论、虚假相关、虚假独立三种现象的产生。